BMC Bioinformatics

BioMed Central

Research article

Open Access

Using quality scores and longer reads improves accuracy of Solexa read mapping

Andrew D Smith†, Zhenyu Xuan† and Michael Q Zhang*

Address: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11274, USA

Email: Andrew D Smith - asmith@cshl.edu; Zhenyu Xuan - xuan@cshl.edu; Michael Q Zhang* - mzhang@cshl.edu

* Corresponding author †Equal contributors

Published: 28 February 2008

BMC Bioinformatics 2008, **9**:128 doi:10.1186/1471-2105-9-128

Received: 5 October 2007

Accepted: 28 February 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/128

Abstract

Background: Second-generation sequencing has the potential to revolutionize genomics and impact all areas of biomedical science. New technologies will make re-sequencing widely available for such applications as identifying genome variations or interrogating the oligonucleotide content of a large sample (*e.g.* ChIP-sequencing). The increase in speed, sensitivity and availability of sequencing technology brings demand for advances in computational technology to perform associated analysis tasks. The Solexa/Illumina 1G sequencer can produce tens of millions of reads, ranging in length from ~25–50 nt, in a single experiment. Accurately mapping the reads back to a reference genome is a critical task in almost all applications. Two sources of information that are often ignored when mapping reads from the Solexa technology are the 3' ends of longer reads, which contain a much higher frequency of sequencing errors, and the base-call quality scores.

Results: To investigate whether these sources of information can be used to improve accuracy when mapping reads, we developed the RMAP tool, which can map reads having a wide range of lengths and allows base-call quality scores to determine which positions in each read are more important when mapping. We applied RMAP to analyze data re-sequenced from two human BAC regions for varying read lengths, and varying criteria for use of quality scores. RMAP is freely available for downloading at http://rulai.cshl.edu/rmap/.

Conclusion: Our results indicate that significant gains in Solexa read mapping performance can be achieved by considering the information in 3' ends of longer reads, and appropriately using the base-call quality scores. The RMAP tool we have developed will enable researchers to effectively exploit this information in targeted re-sequencing projects.

Background

The main technological advances that accompanied the genomic and post-genomic eras are high-throughput sequencing and hybridization microarrays. Sequencing technology enabled scientists to obtain the full genomic sequence for many species, including the human and many model organisms. Sequencing technology is also being used to selectively re-sequence the human genome to detect genome variations such as single nucleotide polymorphisms or large-scale structural variations. Because understanding these variations can immediately impact medical sciences, making sequencing more efficient and accessible is imperative. However, traditional methodologies used to sequence the first mammalian genomes remain expensive, time consuming, and labor intensive.

Oligonucleotide microarray technology, used to interrogate the RNA or DNA content of a sample, has emerged as a widely accessible and effective tool for studying gene expression or detecting protein-DNA interactions (*i.e.* ChIP-chip). Microarrays have also been used to detect genome variations, like SNPs or structural variations. Array-based hybridization also has limitations. For example, probes can behave in highly non-uniform ways, and the effects of cross-hybridization and resolution limits are poorly understood. Although significant research efforts are focused on these problems, they remain inherent in all hybridization-based methods.

The new sequencing technology referred to as "second-generation" shows promise to eliminate many of the problems associated with traditional sequencing technology and also those with oligonucleotide microarray technology. Second-generation sequencers are able to sequence more quickly and at lower cost in terms of both money and labor. New sequencing technologies are developed to sequence at greater depth, meaning that a clone can be sequenced from a sample even when that clone exists at very low abundance (*i.e.* 1 molecule per cell).

Several second-generation technologies have been developed using diverse methods. Two recent second-generation sequencers are from 454 Life Sciences (Roche Diagnostics) and Solexa (Illumina). The 454 sequencers use an emulsion method for DNA amplification and a pyrosequencing protocol for sequencing by synthesis (SBS) at picolitre-scale volumes. Current 454 sequencers can produce 25–50 M nt of sequence in a single run, in the form of reads with length up to 500 nt (a number expected to increase), enabling this technology to be used for *de-novo* sequencing in addition to re-sequencing [1].

Solexa/Illumina 1G sequencers also use sequencing by synthesis, with DNA amplified on the surface of a flow cell, resulting in a random array of dense clusters [2]. The Solexa technology is faster and cheaper than that used in 454 sequencers, producing 1G nucleotides of sequence in one run but producing much shorter reads. Each individual read is roughly 25 to 50 bases in length (also expected to increase slightly in coming years). The Solexa sequencing technology has recently started producing breakthrough results. Research described in [3] and [4] employed Solexa sequencers to obtain high-resolution genomic maps of several histone modifications, as well as localization data for the DNA-binding proteins. Effectiveness of ChIP-sequencing has also been demonstrated by [5], who used Solexa sequencing to obtain locations of STAT1 binding sites in HeLa S3 cells before and following IFN-γ stimulation.

Mapping reads from the Solexa sequencer presents an obvious algorithmic challenge: tens of millions of reads must be mapped to a large (*e.g.* mammalian) genome in a reasonable amount of time. Strong efforts to design short-read mapping algorithms have resulted in methods that are effective in particular contexts. The mapping algorithm implemented as part of the Solexa analysis pipeline is named ELAND (Efficient Large-Scale Alignment of Nucleotide Databases). ELAND is optimized to map very short reads, with length at most 32 nt, and ignores the additional bases when the sequenced reads are longer. ELAND also only allows at most two mismatches between the read and the genomic sequence to which it maps, which will clearly be too few for longer reads. Despite these restrictions, ELAND remains very useful for many mapping tasks because it is extremely efficient. The SXOligoSearch algorithm (by Synamatix) can quickly map reads of varying length, using different criteria, with performance depending on both the read length and mapping criteria. The performance of SXOligoSearch depends on use of the proprietary SynaBASE data structure. This data structure is a heavily compressed and annotated index for the reference genome that retains all non-redundant information. The gains in mapping speed by using such a data structure come at a cost in terms of the memory required for the SynaBASE data structure. Extreme memory requirements of the data structure makes SXOligoSearch unsuitable for use on hardware available in most labs.

In most re-sequencing applications accuracy of mapping is a primary concern. We must know with great accuracy what part of the genome was actually sequenced. There are several reasons why it might be difficult to determine the location in the reference genome from which a read was derived, or even if a read was derived from the reference genome. These include problems with the experiments, such as sequencing errors or sample contamination. Mapping is also made more difficult by repeats in the genome, and by polymorphisms. While mapping algorithms cannot be expected to be robust to all such problems, effort should be made to make the algorithms as robust as possible.

Two sources of information with the potential to improve mapping accuracy are the 3' ends of longer reads, which are often ignored because they contain a higher frequency of errors, and the base-call quality scores. The quality scores describe the confidence of bases in each read. Sequencing quality scores, introduced in the Phred algorithm [6,7], assign a probability to the four possible nucleotides for each sequenced base. The Solexa analysis pipeline, for example, includes a program called BUSTARD to calculate quality scores. Because the bases with lower quality scores are more likely to be sequencing

errors, any potential mapping for a read should be penalized less for mismatching at positions with lower scores. The quality scores are especially important for mapping longer reads, since the 3' ends of longer reads are known to have a higher frequency of sequencing errors.

To investigate whether these sources of information can be used to improve accuracy when mapping reads, we developed the RMAP tool, which can map reads having a wide range of lengths and allows base-call quality scores to determine which positions in each read are more important when mapping. The only requirement on the base-call quality scores for use in RMAP is that they increase monotonically with the inverse of the error probability for a particular base call. Our results indicate that significant gains in mapping performance can be achieved by considering the information in 3' ends of longer reads, and appropriately using the quality scores.

Results and Discussion

The mapping criteria

We designed RMAP to use two different mapping criteria, both based on approximate matching of the read and the reference genome. The first criterion is a simple count of mismatches between a read and the aligned genomic segment. Under this criterion, any unknown nucleotides in the reference genome (*i.e.* Ns) will induce a mismatch with any nucleotide. Uncalled positions, where the sequencing was unable to determine the nucleotide, also induce a mismatch. For a fixed read length, by allowing a greater number of mismatches, more reads can be mapped to reference genome. We refer to this simple mismatch criterion as RMAPM (RMAP using *mismatch* scores).

The second criterion, also based on mismatch-counts, makes use of the base-call quality scores. A cutoff for the base-call quality score is used to designate positions as either high-quality (HQ) or low-quality (LQ), depending on whether the quality score of the highest-scoring base at that position exceeds the cutoff. Low-quality positions always induce a match (*i.e.* act as wild-cards). To prevent the possibility of trivial matches, a quality control step eliminates reads with too many low-quality positions. As with the first criterion, mapping accuracy can be controlled by manipulating the number of allowed mismatches when mapping. But for the second criterion, manipulating the quality-score cutoff provides another means of adjusting sensitivity and specificity, and allows positions to contribute when they are of high-quality, but not be penalized if they are low-quality. We refer to this criterion as RMAPQ (RMAP using *quality* scores).

Evaluating mapping accuracy

Measuring mapping accuracy

In measuring mapping accuracy, we want to quantify both sensitivity and specificity by using reads sequenced from DNA samples from selected genomic regions instead of the entire genome. By mapping those reads to the genome, we can evaluate how accurately they are mapped to the target region. However, there are theoretical and practical limits to how well these can be measured. Inability to map a read correctly can be attributed to sequencing errors (arising from any part of the experiment), to variation between the sampled genome and the reference genome, or can result from ambiguities caused by repeats in the reference genome. These diverse sources of error make it difficult to measure accuracy in terms of traditional sensitivity and specificity.

Ambiguous reads, under a given mapping criterion, are reads that map to more than one location in the reference genome. Reads that map to a single location are called uniquely mappable (or simply mappable) reads. All reads that are not mapped uniquely to some location in the reference genome are said to be unmappable (which includes ambiguous reads). We define target region coverage (or simply coverage) as the number of bases in the target region covered by at least one mappable read, divided by the total number of bases in the target region. In order to compare the coverage values for different read lengths, we use only the first base of each read to represent that read. By counting bases covered, rather than number of reads that map to the target region, greater target region coverage is achieved when the reads map uniformly in the target region. We define mapping *selectivity* as the number of mappable reads that map inside the target region, divided by the total number of mappable reads. A read is said to map inside the target region if any part of the read overlaps the target region. The selectivity shows how well the mapping criterion places mappable reads inside the target region. In this study, when we refer to mapping accuracy, we are referring to both coverage and mapping selectivity (and we formally treat accuracy as the mean of these two measures).

Evaluation data

We used the data from samples of two BACs provided for sequencer validation by Solexa, which covers 162 kb of the chromosome 6 MHC region in an A1-B8-DR3 alternate haplotype assembly based on sequence data from the COX library [8]. We chose one lane of reads sequenced by the 1G sequencers at the CSHL Genome Center. The total number of raw 36 nt reads in this data set is 3.4 million, with the quality score of each base ranging from -5 to 40 (as called by the BUSTARD program from the Solexa analysis pipeline). As reference genome we used hg18, all chromosomes except chr6, which we replaced entirely

with chr6_cox_hap1, the A1-B8-DR3 alternate haplotype assembly. We chose to include this alternate haplotype in the reference because it is the origin of the BAC that was sequenced. We excluded the ordinary chr6 because it has high similarity with chr6_cox_hap1, and including both of these would have resulted in a high proportion of reads mapping ambiguously to chr6 and chr6_cox_hap1 (see additional file 1).

Evaluation procedure

To investigate how information is distributed within the reads, we ran RMAP on all reads with lengths ranging from 25–36 nt, allowing mismatches in the range of 0 to 10, and using both the RMAPM and RMAPQ criteria. For the RMAPQ criterion, we chose {4, 8, 12, 16, 20, 24} as the set of quality-score cutoffs to evaluate. Reads with fewer than 10 contiguous HQ bases (*i.e.* bases scoring above the quality-score cutoff) were considered unmappable and removed from consideration, as the algorithm requires a minimum number of high-quality bases for efficiency (see Methods section for details). Any read lacking 10 consecutive high-quality bases would likely have a very high overall amount of error.

Mapping longer reads with more mismatches increases accuracy

The Solexa sequencer can produce reads of more than 50 bases, and longer reads contain more sequence information. Although it is known that the quality of sequenced bases in reads decreases toward the 3' end of the read, especially as read length increases, it remains to be shown how much useful information may still exist in bases at the 3' ends of longer reads. Making use of any additional bases is only expected to improve mapping accuracy if the additional bases contain information of sufficient quality. When the BAC reads were mapped to the human genome using the RMAPM criterion, with length from 25–36 nt, and different number of allowed mismatches, we found that there is generally a great deal of information in 3' end bases up to 36 nt. These results are presented in Figure 1 and Supplementary Table 1 (see additional file 2) (We remark that the accuracy of RMAP using the RMAPM criterion for reads shorter than 32 nt, allowing at most 2 mismatches, is the same as that of ELAND). The BAC coverage always increased with length of mapped reads, except when only one or zero mismatches are allowed. The mapping selectivity decreases monotonically with read length when zero or at most one mismatch is allowed. When multiple mismatches are allowed, the mapping selectivity first increases with read length, then decreases slightly. Taking the mean of these two measures as overall mapping accuracy, we see that the combined target region coverage and mapping selectivity is maximized when read length is 36 nt and up to 4 mismatches are allowed. Comparing read lengths between 25 nt and 36 nt, the mapping

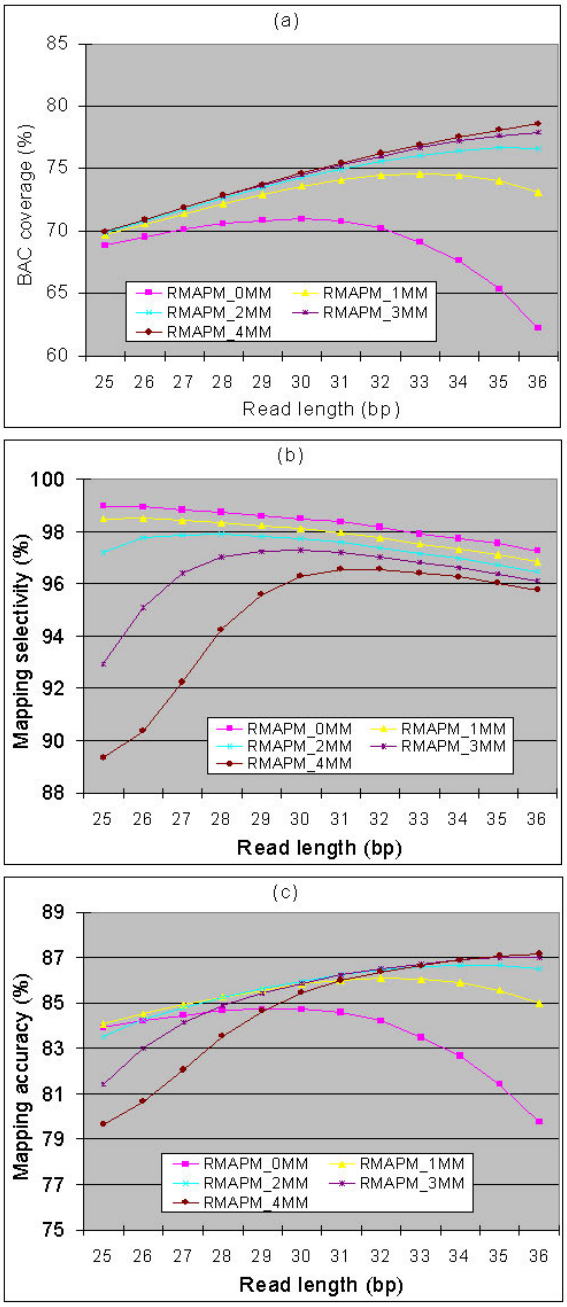

Figure 1
Comparison of mapping accuracy of RMAPM criterion under different parameter combinations. Comparison of mapping accuracy for reads of different lengths, and allowing different numbers of mismatches without using quality scores. Both the target (BAC) region coverage (a) and the mapping selectivity (b) are displayed. The mean of these two measures is presented in (c) as mapping accuracy. Standard error of displayed values was always ≤ 1.0% and usually < 0.1%, as estimated by mapping reads obtained from the second lane of the same sequencing run of the same BAC regions (this applies also to values in Figure 2).

selectivity increases 6.4% while the BAC region coverage increases 8.6%. By extrapolating our results, even greater improvements are expected as read lengths increase beyond 36 nt.

Using quality score information increases accuracy

We tested different cutoffs for defining high quality bases to estimate the ability of the RMAPQ criterion to most effectively use the quality information in the reads. We achieve better mapping performance in both BAC coverage and mapping selectivity when using longer reads with high quality score cutoff of 4 to 24 than without using quality score filtering (See Figure 2, and additional file 2). This is due to a larger number of reads being mapped unambiguously to the genome by the RMAPQ criterion, and a larger number of those being mapped to the target (BAC) regions. The best mapping accuracy was achieved when reads were of length 36 nt, a quality-score cutoff of 8 was used, and when at most one mismatch was allowed. For these settings, the mapping selectivity further improved almost 3% with the same BAC coverage, compared with the best accuracy of RMAP using the RMAPM criterion. These results demonstrate that the way in which the quality scores are incorporated into RMAP results in improved accuracy.

The reads derived from the BACs provided by Solexa for the purpose of validation are of very high quality. For a given 162 Kb BAC region, most of these 3.4 million reads can be mapped almost perfectly. This introduces a saturation effect, leaving limited space for improvement, as highly accurate mapping is achieved easily. RMAP still makes significant improvements within this narrow range. In many applications, without this ceiling effect, the improvement is expected to be even more pronounced.

We also compared the mapping accuracy of RMAP with another available, but presently unpublished, method called MAQ [9]. Using default parameter setting, we ran MAQ to map the set of 3.4 M reads from the BAC. Although MAQ had speed and memory usage similar to RMAP, we found MAQ to have lower mapping accuracy (see additional file 3) when allowing at most 2 mismatches.

Discussion

Widely-accessible second-generation sequencing technologies promise to revolutionize many areas of bio-medical research. In addition to *de novo* sequencing of new species, these technologies make targeted re-sequencing a reality. Re-sequencing will provide more accurate means of interrogating the oligo-nucleotide content of samples, and of identifying important genome variations, such as disease-related SNPs. Mapping reads to genomes is a critical step in re-sequencing data analysis, and both the algorithmic and software technology must keep pace with surging advances in the throughput of sequencing instruments.

In order to maximize the use of available information in mapping Solexa reads, we developed the RMAP tool, which incorporates base-call quality scores to improve accuracy. RMAP responds to an urgent need for such an algorithm by providing both the accuracy to handle emerging mapping tasks. Our results in applying RMAP have shown that more reads can be mapped into the target regions when using the RMAPM criterion to map longer reads and allow more mismatches. We have also shown that the way in which quality scores are used in RMAP (the RMAPQ criterion) significantly increases both coverage and mapping selectivity.

Although second-generation sequencing technology is currently producing many important results, there is still little understanding of how this technology should behave with respect to sequencing errors and what are the general properties of typical re-sequencing data sets. As more knowledge accumulates about typical results from this new sequencing technology, more information can be incorporated into algorithms for mapping reads and other associated analysis tasks.

In theory we could move toward an ideal mapping criterion by predictive modeling, where a model would be trained to identify the location from which each read was derived. Although the best mapping criteria may not be amenable to high-throughput computation, some approximation of those criteria could be developed. Cross-validation and the use of a wide range of data sets could be used to ensure that the trained criteria are sufficiently general. In practice such a procedure would require high-quality training data, and an extreme amount of computing time to train and evaluate such models.

RMAP does not consider insertions or deletions (indels), which are potentially important in certain sequencing applications (*e.g.* indel polymorphisms). The straight-forward strategy for handling indels is to extend initial seed matches using a Smith-Waterman-style alignment, as is commonly done in database search programs like BLAST [10]. For short reads, with length ≤ 50 bases, providing this greater flexibility will require careful investigation into scoring the alignments, because using simple scoring schemes for indels may result in higher rate of false-positive mappings and ambiguities. In addition, because indels have nothing analogous to the base-call quality scores, it will be more difficult to distinguish errors from real genotypic variations during the mapping stage of analysis. We have observed that data sets containing a high proportion of low-complexity reads, such as single

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.